ISS-A-066

自然言語処理によるデータ分析の自動化に関する研究

佐々木　慎太朗†　　井上　浩孝††

† 呉工業高等専門学校専攻科プロジェクトデザイン工学専攻　　†† 呉工業高等専門学校電気情報工学科

1. はじめに

会社などでの仕事の中にはマーケティングの問題などを示すデータ分析などの業務を使う職種は多くある．例えば，顧客に対してのサービスや物の売り上げの見直しなどがそうだ．そして，そのデータ分析の過程の作業では手間と時間を浪費する．実際にエクセルを用いて手動でデータ分析を行ってみるとデータの整理，前処理の段階で多くの時間を要した。そこで，その作業時間を短縮できないかと考えた．本研究では，これを目的として，自然言語処理の手法を一部用いて作業補助及び，自動化を試みた．自然言語処理は自然言語をコンピューターに処理させる技術の総称で様々な言語モデル，手法が提案されている．

2. 分析手順

参考文献[1]を元に実際にデータ分析を行った．その際の手順を元に自動化を行う．次にそこで必要な作業を記述する．最初にデータを収集するが，この際に扱ったのは kickstarter というサイトである[2]．このサイトでは膨大な情報が集積されていて様々なグラフがとれて試験的に分析しやすい．大抵その収集されたデータは自動的に収集されたデータでありひとまとめにして管理され，それを分析する際にはエクセルなどを使って分析しやすい形にする前処理の作業が必要となる．その次に前処理をしたテーブルを元に様々なグラフを作成し視覚化した情報から分析を行う．今回収集したデータから行った分析の作業では図 1 のようなグラフがとれた．

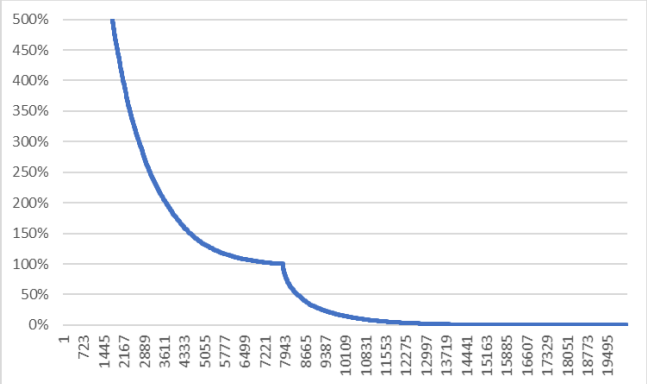

図1. 達成率と目標金額

前述したサイトは，利用者がプロジェクトを立ち上げ資金を募り設けられた期間内にその額に達することを目的としている．このグラフはプロジェクトの達成度合いと達成する為に必要な目標金額の相関を調べるために作成した．縦軸がプロジェクト達成率，横軸が設定された目標金額となっている．達成率が 100%以上なら成功していてそれ以下なら失敗している．このグラフを見ると目標金額が低いほど達成しやすくなっていて高いほど失敗しっやすいことが分かる．そして，100%付近を見ると特異点が見られる．つまり，達成率に関してのグラフと合わせて見ると何か分かるかもしれないと推測できる．終了したか否かで調べると特異点は無く，プロジェクト終了間際に現れるものだということが分かる．この結果，達成しそうなプロジェクトがあるとそれを目当てに支援する人が急増するなどが考えられる．ここまでが分析の一連の流れである．このような前処理，グラフ作成，分析の一連の流れを自動化，作業補助をする．

3. 自動化

現段階では前処理の段階を自動化している．Python を使ってエクセルを操作して，膨大な生のデータの塊であるテーブルから必要な列や行を取り出して，新たなグラフ作成用のテーブルを作成する．グラフ作成ではエクセルを操作する関数を使って様々なグラフを自動的に生成させることを検討している．そして，最後の分析作業では自然言語処理の手法により文章生成をして一部補助，自動化をする．

4. 今後の課題

文章生成では分析対象のグラフを元に，生成するより良い特徴量を調査しながら RNN などの言語モデルを使ってどのようなグラフか言及する文章を生成させることを検討している．

参考文献

[1]　有賀康顕，中山心太,西林孝，「仕事ではじめる機械学習」，オライリージャパン，2018.

[2]　https://www.kickstarter.com/